CLAIMS

WHAT IS CLAIMED IS:

1. A method for determining a genotype of at least one individual from a genetic marker using at least one measure of the amount of a given allele of the genetic marker in the individual, comprising:

assigning the measure of the amount of the allele to a group using one or more of a probability clustering process and a distance-based clustering process; and

assigning a genotype to the group based on a property of the group, wherein the individual is determined to have the genotype assigned to the group.

2. A method as claimed in claim 1, wherein the method is computer-implemented.

3. A method as claimed in claim 1, wherein the genetic marker is a SNP position.

4. A method as claimed in claim 1, wherein the individual is a haploid organism.

5. A method as claimed in claim 1, wherein the individual is a diploid organism.

6. A method as claimed in claim 5, wherein the diploid organism is a mammal.

7. A method as claimed in claim 6, wherein the mammal is a human.

8. A method as claimed in claim 1, wherein the probability clustering process carries out an expectation maximization algorithm.

9. A method as claimed in claim 1, wherein the distance-based clustering process carries out at least one K-means algorithm.

10. A method as claimed in claim 9, wherein the at least one K-means algorithm is initiated by assigning a plurality of mean values evenly distributed between approximately 0 and approximately 1.

11. A method as claimed in claim 10, wherein 10 mean values are assigned.

12. A method as claimed in claim 9, wherein the at least one K-means algorithm determines a solution for a plurality of subsets of density centers.

13. A method as claimed in claim 12, wherein each subset comprises three density center values.

14. A method as claimed in claim 12, wherein each subset comprises two density center values.

15. A method as claimed in claim 12, wherein each subset comprises one density center value.

16. A method as claimed in claim 12, wherein the plurality of subsets of density center values comprises every combination of subsets of density center values.

17. A method as claimed in claim 9, comprising carrying out a first K-means algorithm and a second K-means algorithm.

18. A method as claimed in claim 17, comprising carrying out a first K-means algorithm and a second K-means algorithm, wherein the first K-means algorithm is initiated by assigning a plurality of mean values; and the second K-means algorithm determines a solution for a plurality of subsets of density centers obtained by the first K-means algorithm.

19. A method as claimed in claim 1, wherein the probability clustering process is initiated using a seed value.

20. A method as claimed in claim 1, wherein the probability clustering process is initiated using a solution obtained by at least one K-means algorithm.

21. A method as claimed in claim 1, wherein a solution obtained by the probability clustering process and/or the distance-based clustering process yields a minimum maximum standard deviation for a distribution of the at least one measure of the amount of the allele.

22. A method as claimed in claim 1, comprising assigning the measure of the amount of the allele using both a probability clustering process and a distance-based clustering process.

23. A method as claimed in claim 22, wherein the probability clustering process carries out an expectation maximization algorithm and the distance-based clustering process carries out a K-means algorithm.

24. A method as claimed in claim 23, wherein the measure of the amount of the allele is assigned to a group based on a comparison between a) a solution determined by the K-means algorithm and b) a solution determined by the expectation-maximization algorithm, wherein the measure of the amount of the allele is assigned to a group according to the solution yielding a minimum maximum standard deviation.

25. A method as claimed in claim 1, wherein the property of the group is a characterizing property of the group and the genotype is assigned based on the characterizing property of the group falling within a one of a plurality of ranges of values of the measure of the amount of the allele, each of the plurality of ranges of values corresponding to a different genotype.

26. A method as claimed in claim 1, wherein the method determines the genotype of a plurality of individuals using a plurality of respective measures of the amount of the allele of the same genetic marker in each of the individuals, and wherein each of the measures of the amount of the allele is assigned to a one of a plurality of potential groups.

27. A method as claimed in claim 26, further comprising;

assessing a confidence of the determinations of the genotypes of the plurality of individuals based on a criterion for p-values corresponding to a particular confidence level.

28. A method as claimed in claim 27, wherein assessing the confidence includes carrying out at least a first and a second evaluation of the confidence in the determinations.

29. A method as claimed in claim 28, wherein the first evaluation comprises a chi-squared distribution to determine the confidence in the determinations.

30. A method as claimed in claim 29, wherein a chi-squared distribution p-value is calculated for each standard deviation of each group from a chi-squared distribution based on a pre-set maximum standard deviation cutoff and a number of degrees of freedom reflecting the number of genotypes in a given group, and the determination of genotypes is rejected if one of the chi-squared distribution p-values does not meet a criterion corresponding to a confidence level.

31. A method as claimed in claim 30, wherein the maximum standard deviation cutoff is set to a value of 0.05.

32. A method as claimed in claim 30, wherein the confidence level is 99.9%.

33. A method as claimed in claim 28, wherein the second evaluation comprises determining a likelihood of the assigned distributions conforming to a corresponding Hardy-Weinberg equilibrium for the plurality of individuals.

34. A method as claimed in claim 33, wherein determining the likelihood includes:

calculating a first sum of a first set of Bayesian factors for all possible permutations of the plurality of individuals over the plurality of potential groups;

calculating a second sum-of Bayesian factors from a lowest Bayesian factor in the first set to the Bayesian factor corresponding to the assigned distribution of the individuals between the groups; and

determining a p-value from the quotient of the second sum and the first sum.

35. A method as claimed in claim 34, wherein the determination of genotypes is rejected if the p-value does not meet a criterion corresponding to a confidence level.

36. A method as claimed in claim 35, wherein the confidence level is 99.9%.

37. A method as claimed in claim 28, further comprising determining a ratio between a likelihood of a measure of the amount of an allele corresponding to the assigned group and a likelihood of the measure of the amount of an allele corresponding to the next best fit group.

38. A method as claimed in claim 26, wherein at least one solution from the probability clustering process and the distance-based clustering process includes validating a number of groups to which the measures of the amount of the allele are to be assigned.

39. A method as claimed in claim 38, wherein validating the number of groups includes assigning the measures of the amount of the allele to a first number of groups and using a property of the groups to determine whether the assignment to the first number of groups is reliable.

40. A method as claimed in claim 39, wherein the property of the groups is a mean or median value of each group.

41. A method as claimed in claim 40, wherein using a property of the groups includes determining whether the mean or median values of the groups are sufficiently dissimilar for the groups to constitute different groups.

42. A method as claimed in claim 41, wherein the mean or median values of the groups are considered to be sufficiently dissimilar if they differ by more than a cut off corresponding to a difference which minimizes a number of incorrect genotype assignments for two independent genotype assignments for the same individuals.

43. A method as claimed in claim 39, wherein if it is determined that the assignment to the first number of groups is reliable, then genotypes are assigned to the groups.

44. A method as claimed in claim 39, wherein if it is determined that the assignment to the first number of groups is unreliable, then one or more of the probability clustering process and the distance-based clustering process is repeated to assign the measures of the amount of the allele to a second number of groups different from the first number of groups.

45. A method as claimed in claim 44, wherein the second number of groups is less than the first number of groups.

46. A method as claimed in claim 44, wherein if the measures are assigned to three groups, then the genotype is assigned depending on a ranking of the groups, and if the measures are assigned to less than three groups, then the genotype is assigned depending on a mean value of each group.

47. A method as claimed in claim 46, wherein the genotype is selected from the group consisting of: homozygous reference; homozygous alternate; and heterozygous.

48. A method as claimed in claim 43, wherein the K-means clustering process includes determining a representative value for each group and assigning a genotype to each group based on the respective representative values of each group.

49. A method as claimed in claim 48, wherein assigning the genotype includes determining whether the representative value of a group falls within a one of a plurality of ranges of values.

50. A method as claimed in claim 48 or 49, wherein the representative value is a mean or median of a group.

51. A method as claimed in claim 49, wherein there are three ranges of values and a first range corresponds to a homozygous reference, a second range corresponds to a heterozygous and a third range corresponds to a homozygous alternate.

52. A method as claimed in claim 49, wherein the plurality of ranges of values have been determined by calibrating the measure of the amount of the allele using a sufficiently large sample of individuals to allow groups corresponding to all the different genotypes to be unambiguously determined.

53. A method as claimed in claim 52, wherein at least one boundary of each range is the value at which adjacent corresponding groups intersect.

54. A method as claimed in claim 44, wherein the probability clustering process uses a first set of seed values to assign the first number of groups and a second set of seed values to assign the second number of groups.

55. A method as claimed in claim 54, wherein the seed values are means for respective groups.

56. A method as claimed in claim 55, wherein the first set of seed values comprises approximately 0.2, 0.5 and 0.8, and the second set of seed values comprises approximately 0.3 and 0.7.

57. A method as claimed in claim 55, wherein a ratio of the members of the first set of seed values is approximately 1:1:1 and a ratio of the members of the second set of seed values is approximately 1:1.

58. A computer implemented method for determining one or more genotypes of a plurality of individuals at a SNP position using respective measures of a relative allele amount for the SNP position for each individual, comprising:

assigning the measures of the relative allele amount to a group using one or more of an expectation maximization process and a K-means process;

assigning a genotype to each group identified by the expectation maximization process and/or the K-means process to determine a genotype of each person; and

assessing a confidence of determination of the genotype.

59. A method as claimed in claim 58, wherein the expectation maximization process is initiated using a K-Means algorithm.

60. A method as claimed in claim 58, wherein assessing the confidence includes using a chi-squared distribution to evaluate a spread of at least one of said groups and evaluating whether a distribution of the individuals between the groups conforms to a corresponding Hardy-Weinberg equilibrium distribution.

61. A method as claimed in claim 58, wherein the expectation maximization process determines a number of groups to which to assign the measures and assigns the measures to less than three groups if it is determined that an assignment to three groups would be unreliable.

62. A method as claimed in claim 61, wherein if the measures are assigned to three groups, then the genotype is assigned depending on the rank of the groups, and if the measures are assigned to less than three groups, then the genotype is assigned depending on a mean value of each group.

63. A data processing apparatus for determining the genotype of at least one individual from a genetic marker using at least one measure of the amount of a given allele of the genetic marker in the individual, comprising:

a data processor;

a storage device holding computer readable code in communication with the data processor, the computer readable code including:

computer code which assigns the measure of the amount of the allele to a group by executing one or more of a probability clustering process and a distance-based clustering process; and

computer code which assigns a genotype to the group based on a property of the group and determines the individual to have the genotype assigned to the group.

64. A data processing apparatus as claimed in claim 63, wherein the computer code executing a probability clustering process carries out an expectation maximization algorithm.

65. A data processing apparatus as claimed in claim 63, wherein the computer code executing a distance-based clustering process carries out at least one K-means algorithm.

66. A data processing apparatus as claimed in claim 63, wherein the probability clustering process is initiated using a seed value.

67. A data processing apparatus as claimed in claim 63, wherein the probability clustering process is initiated using a solution obtained by at least one K-means algorithm.

68. A data processing apparatus as claimed in claim 63, wherein a solution obtained by the probability clustering process and/or the distance-based clustering process yields a minimum maximum standard deviation for a distribution of the at least one measure of the amount of the allele.

69. A data processing apparatus as claimed in claim 67, wherein the computer code executes both a probability clustering process and a distance-based clustering process.

70. A data processing apparatus as claimed in claim 69, wherein the probability clustering process carries out an expectation maximization algorithm and the distance-based clustering process carries out at least one K-means algorithm.

71. A data processing apparatus as claimed in claim 70, wherein the computer code assigns the measure of the amount of the allele by comparing a) a solution determined by the K-means algorithm and b) a solution determined by the expectation-maximization solution, wherein the measure of the amount of the allele is assigned to a group according to the solution yielding the minimum maximum standard deviation.

72. A data processing apparatus as claimed in claim 63, further comprising computer code for determining a confidence of the determination of genotype of at least one individual.

73. A computer readable medium comprising computer readable code for determining the genotype of at least one individual from a genetic marker using at least one measure of the amount of an allele of the genetic marker in the individual, and for carrying out the processes of:

assigning the measure of the amount of an allele to a group using one or more of a probability clustering process and a distance-based clustering process; and

assigning a genotype to the group based on a property of the group and determining the individual to have the genotype assigned to the group.

74. A computer readable medium as claimed in claim 73, wherein the probability clustering process comprises an expectation maximization algorithm.

75. A computer readable medium as claimed in claim 73, wherein the distance-based clustering process comprises a K-means algorithm.

76. A computer readable medium as claimed in claim 73, wherein the measure of the amount of the allele is assigned using both a probability clustering process and a distance-based clustering process.

77. A computer readable medium as claimed in claim 73, further comprising computer readable code for determining the confidence of the determination of genotype of at least one individual.

78. A method for calibrating a range of values of measures of the presence of an allele at a genetic marker, comprising:

measuring the value of a measure of the presence of an allele for sufficient individuals to allow groups of individuals corresponding to each expected genotype to be determined, wherein each expected genotype consists of: homozygous reference, homozygous alternate and heterozygous;

determining boundaries between adjacent distributions of values of the measure for each group of individuals; and assigning a range of values of the measure of the presence of the allele to each genotype, wherein at least one boundary of each range is determined by an intersect of the distributions of adjacent groups.

79. A method for establishing a cut off separation in groups of values of measures of a presence of a given allele at a genetic marker corresponding to different groups for the purposes of carrying out a probability clustering process, comprising:

measuring the values of the presence of the allele at the genetic marker at least twice for the same group of individuals;

assigning the values of the presence of the allele to a number of groups using a clustering algorithm in which the number of groups to which to assign the values is determined using a cut off separation of assigned groups for the independent measurements;

for a cut off value, assigning a genotype to each group for each of the individuals and for each of the independent measurements; and

minimizing a number of inconsistent genotype assignments for the same individual as a function of the cut off.